AA 274
Principles of Robotic Autonomy
Stereo vision and structure from motion
Logistics

• It’s the final (project) stretch!
  • All sections are open office hours for project discussion with TAs
  • Final project check-in was due Tuesday
  • Final project demos: Thursday, December 15\textsuperscript{th}, 3:30 – 6:30pm
  • Sign up for the correct slots with correct group number
Today’s lecture

• Aim
  • Learn fundamental geometric concepts needed for 3D reconstruction
  • Learn basic techniques to recover scene structure, chiefly stereo and structure from motion

• Readings
  • SNS: 4.2.5 – 4.2.7
The see-think-act cycle

Knowledge

Localization
Map Building

environmental model
local map

Information extraction

raw data

Sensing

Decision making
Motion planning

position
global map

trajectory

Trajectory execution

actuator commands

Actuation

Mission goals

Real world environment

See-think-act
Measuring depth

Once the camera is calibrated, can we measure the location of a point $P$ in 3D given its known observation $p$?

- No: one can only say that $P$ is located somewhere along the line joining $p$ and $O$!
Recovering structure

• **Structure**: 3D scene to be reconstructed by having access to 2D images

• **Common methods**
  1. Through recognition of landmarks (e.g., orthogonal walls)
  2. Depth from focus: determines distance to one point by taking multiple images with better and better focus
  3. Stereo vision: processes two distinct images taken at the *same time* and assumes that the relative pose between the two cameras is *known*
  4. Structure from motion: processes two images taken with the same or different cameras at *different times* and from different *unknown* positions
Stereopsis (why we have two eyes)
Binocular reconstruction

- **Given:** calibrated stereo rig and two image matching points $p$ and $p'$
- **Find** corresponding scene point by intersecting the two rays $\overline{Op}$ and $\overline{O'p'}$ (process known as **triangulation**)

![Diagram showing binocular reconstruction with points O, P, O', P']
Approximate triangulation

- Due to noise, triangulation problem is often solved as finding the point $Q$ with images $q$ and $q'$ that minimizes

$$d^2(p, q) + d^2(p', q')$$

Re-projection error
Stereo vision process

• Stereo vision consists of two steps:
  1. *fusion* of features observed by two (or more) cameras -> correspondence
  2. *reconstruction* of their three-dimensional preimages -> triangulation

• Step 2 is relatively easy; Step 1 requires you to establish correct correspondences and avoid erroneous depth measurements

• Several constraints can be leveraged to simplify Step 1 (e.g., similarity constraint, continuity constraints, etc.); most important: epipolar constraint
Consider images $p$ and $p'$ of a point $P$ observed by two cameras from $O, O'$.

These five points all belong to the **epipolar plane** defined by $p, O, O'$, or equivalently, $p', O, O'$.

**Epipolar constraint**: potential matches for $p$ must lie on epipolar line $l'$ (and vice-versa).
Epipolar constraint

• Search for matches can be restricted to the epipolar line instead of the whole image! → one dimensional search
Epipolar constraint: derivation

- Epipolar constraint: $OP, O'P'$, and $OO'$ must be coplanar, or

$$\overrightarrow{OP} \cdot \left[ \overrightarrow{OO'} \times \overrightarrow{O'P'} \right] = 0$$
Aside: matrix notation for cross product

- Cross product can be expressed as the product of a skew-symmetric matrix and a vector

\[
a \times b = \begin{bmatrix}
0 & -a_3 & a_2 \\
a_3 & 0 & -a_1 \\
-a_2 & a_1 & 0
\end{bmatrix}
\begin{bmatrix}
b_1 \\
b_2 \\
b_3
\end{bmatrix}
= [a] \times b
\]
Epipolar constraint: derivation.  
More details: CS231a

- Assume that the world reference system is co-located with camera 1
- After some algebra, epipolar constraint becomes [FP, Section 7.1]

$$p^T F p' = 0$$

where: $$F = K^{-T} [t] \times R K'^{-1}$$
Key facts

• $F$ is referred to as the fundamental matrix

• $l = Fp'$ (resp. $l' = F^T p$) represents the epipolar line corresponding to the point $p'$ (resp. $p$) in the first (resp. second) image. This exploits the homogenous notation for lines.

• $F^T e = Fe' = 0 \Rightarrow F$ is also singular (as $t$ is parallel to the coordinate vectors of the epipoles)

• $F$ has 7 DoF (9 elements – common scaling – $\det(F)=0$)
Usefulness of fundamental matrix

• Assume $F$ is given
• Given a point in image 1, one can compute the corresponding epipolar line in image 2 without any additional information needed!
Estimating the fundamental matrix

• 8-point algorithm

\[ p = [u, v, 1]^T, \quad p' = [u', v', 1]^T \quad \Rightarrow \quad \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = 0 \]

\[ \Rightarrow \begin{bmatrix} uu' & uv' & u & vu' & vv' & v & u' & v' & 1 \end{bmatrix} \begin{bmatrix} F_{11} \\ F_{12} \\ F_{13} \\ F_{21} \\ F_{22} \\ F_{23} \\ F_{31} \\ F_{32} \\ F_{33} \end{bmatrix} = 0 \quad \Rightarrow \quad Wf = 0 \]

• Given \( n \geq 8 \) correspondences, one then solves

\[
\min_{f \in \mathbb{R}^9} \| Wf \|^2 \quad \text{subject to} \quad \|f\|^2 = 1
\]

\[ \Rightarrow \tilde{F} \]
Enforcing the rank constraint

- \( \tilde{F} \) satisfies the epipolar constraints, but is not necessarily singular (hence, is not necessarily a proper fundamental matrix)
- Enforce rank constraint (again, via SVD decomposition)

\[
\text{Find } F \text{ that minimizes } \| F - \tilde{F} \|^2 \quad \text{subject to } \det(F) = 0
\]

- 8-point algorithm
  1. Use linear least squares to compute \( \tilde{F} \)
  2. Enforce rank-2 constraint via SVD
Parallel image planes

- Assume image planes are parallel
- Epipolar lines are horizontal
- \( v \) coordinates are equal
  - Easier triangulation
  - Easier correspondence problem
- Is it possible to warp images to simulate a parallel image plane?
Image rectification

• Achieved by applying an appropriate projective transformation
• Several algorithms exist
• From now on, we assume rectified image pairs
Back to stereo vision process

• Recall that stereo vision consists of two steps:
  1. fusion of features observed by two (or more) cameras (correspondence)
  2. reconstruction of their three-dimensional preimages (triangulation)

• Correspondence problem

Goal: find corresponding observations \( p \) and \( p' \)
Exploits epipolar constraints
Two classes of algos: area-based and feature-based
Hard problem: occlusions, repetitive patterns, etc.; more on this later
Triangulation under rectified images

• We already saw how to triangulate correspondences in the general case

• **Triangulation problem** under rectified images:

From similar triangles:

\[ z = \frac{bf}{p_u - p'_u} \]

Large baseline: Object might be visible from one camera, but not the other

Small baseline: large depth error
Disparity map

- Disparity: pixel displacement between corresponding points
- Disparity map: holds the disparity values for every pixel
- Nearby objects experience largest disparity
Method #3: structure from motion (SFM)

Given $m$ images of $n$ fixed 3D points

$$p_{j,k} = M_k P_j$$

Find:
- $m$ projection matrices $M_k$ (motion)
- $n$ 3D points $P_j$ (structure)
SFM ambiguity

• It is not possible to recover the absolute scale of the observed scene
Solution to SFM problem (high-level)

• Several approaches available:
  • Algebraic approach (by fundamental matrix)
  • Bundle adjustment

• Algebraic approach (2-views)
  1. Compute fundamental matrix $F$ (e.g., via 8-point algorithm)
  2. Use $F$ to estimate projection camera matrices
  3. Use projection camera matrices for triangulation
Application of SFM: visual odometry

- **Visual odometry**: estimate the motion of the robot by using visual input (and possibly additional information)
  - Single camera: absolute scale must be estimated in other ways
  - Stereo camera: measurements are directly provided in absolute scale
Thanks for a great quarter!